JOURNAL OF NUMBER THEORY **67**, 146–161 (1997)
ARTICLE NO. NT972175

# Transcendence of Numbers with a Low Complexity Expansion

Sébastien Ferenczi* and Christian Mauduit†

*Institut de Mathématiques de Luminy, CNRS—UPR 9016, Case 930-163 avenue de Luminy, F13288 Marseille Cedex 9, France*

*Communicated by M. Waldschmidt*

Received January 3, 1996; revised April 15, 1997

A sequence is Sturmian if it has complexity $n+l-1$, that is, $n+l-1$ factors of



particularly the Arnoux–Rauzy sequences. © 1997 Academic Press

It is well known that, for any integer $k \geqslant 2$, the expansion in base $k$ of a rational number is ultimately periodic. If we decide to measure the "complexity" of a $k$-adic expansion by counting the number of blocks of digits of length $n$ which appear in it, we may then say that the complexity of the $k$-adic expansion of a rational number is very low. In this work, we consider the following problem: can there exist irrational algebraic numbers whose expansion in some base $k$ is of "low complexity?" (This low complexity represents in a way the opposite situation to normal numbers in base $k$.)

Our main result shows that if the $k$-adic expansion of an irrational number has the lowest possible complexity (in that case, we say it is a *Sturmian* real number, after the definition of Hedlund and Morse [HED-MOR]), then this number is transcendental.

This result was known only in some particular cases ([ADA-DAV], [DAV], [KOM] correcting [NIS-SHI-TAM]; see the discussion at the beginning of Section 4). The method we propose uses a combinatorial translation of a result of Ridout [MAH, Chap. 9, pp. 147–148] which may be of independent interest (Proposition 1), stating that, if the expansion of a number contains infinitely many $(2+\varepsilon)$-powers of blocks (that is, a block followed by itself and then by its beginning of relative length at least $\varepsilon$), at

* E-mail: ferenczi@iml.univ-mrs.fr.
† E-mail: mauduit@iml.univ-mrs.fr.



0022-314X/97 $25.00

distances from the origin which are not too much larger than the lengths of the considered blocks, then it is transcendental. The transcendence in the Sturmian case is then a consequence of this criterion and of the combinatorial properties of Sturmian sequences; if we add to this the result of Hedlund and Morse that Sturmian sequences are produced by rotations, we get also an algebraic expression of the numbers with Sturmian expansion, either in base 2 (the classical case of Hedlund and Morse), or in any base $k$, using a new characterization of Sturmian sequences on $k$ letters.

The same method allows us also to consider other classes of real numbers with a low complexity $k$-adic expansion, in particular the Arnoux–Rauzy sequences, which form a subclass of the sequences of complexity $2n+1$. It applies also naturally to real numbers whose expansion is given by some recursion formulas (e.g. Chacon's sequence) or by some automatic rules, which allows us to give a new proof of some of the results in [LOX-vdP] on the transcendence of a special class of real numbers with automatic expansions.

We thank the referee for his valuable comments and suggestions.

## 1. DEFINITIONS

For a finite alphabet $A$, we call a finite string of elements in $A$ a *word* (or *block*); the set of all finite nonempty words on $A$ is denoted by $A^*$.

The *concatenation* of two words $v = v_1 \cdots v_r$ and $w = w_1 \cdots w_s$ is the word $vw = v_1 \cdots v_r w_1 \cdots w_s$. In the same way, we can concatenate a word with a (one-sided) infinite sequence $u_0 u_1 \cdots$.

*Prefixes* and *suffixes* are defined as usual.

*Sequences* on the alphabet $A$ are one-sided and denoted by $u = (u_n\ n \in \mathbb{N}) = u_0 u_1 \cdots u_n \cdots$.

A word $v = v_1 \cdots v_r$ is said to *occur* in or to be a *factor* of the sequence $u$ if there exists an $m$ such that $u_m = v_1, ..., u_{m+r-1} = v_r$.

The sequence $u$ is *recurrent* when every factor of $u$ occurs infinitely often in $u$. The sequence $u$ is *minimal* when every factor of $u$ occurs infinitely often in $u$ and with bounded gaps.

The *language* $L_n(u)$ is the set of all factors of length $n$ of $u$. The *language of the sequence* is the reunion of all the $L_n(u)$.

The *complexity function* of the sequence $u$ is the function associating to $n$ the cardinality of $L_n(u)$, denoted by $p(n)$.

In this paper, a *substitution* is a map $\sigma$ from an alphabet $B$ to the set $A^*$; it extends to a morphism (for the concatenation) from $B^*$ to $A^*$ by $\sigma(v_1 \cdots v_r) = \sigma(v_1) \cdots \sigma(v_r)$ for any word $v_1 \cdots v_r$ in $A^r$; it can also be applied to an infinite sequence by defining $\sigma(u_1 \cdots u_n \cdots)$ to be the infinite concatenation of the $\sigma u_i$.

A substitution from $A$ to $A^*$ is more simply called a substitution on $A$; it is called *primitive* if there exists $k$ such that $a$ occurs in $\sigma^k b$ for any $a \in A$, $b \in A$. It is called *of constant length* $q$ if $\sigma a$ is of length $q$ for any $a \in A$.

A *fixed point* of $\sigma$ is an infinite sequence $u$ with $\sigma u = u$.

We recall a famous result of Hedlund and Morse ([HED-MOR]): if there exists $n$ such that $p(n+1) = p(n)$, or such that $p(n) \leqslant n$, then $u$ is ultimately periodic.

We are looking at non-ultimately periodic sequences with the lowest possible complexity. The sequence $u$ is called *Sturmian* (on two letters) if $p(n) = n+1$ for every $n \in \mathbb{N} \setminus \{0\}$, or, more generally, for any integer $l \geqslant 2$, the sequence $u$ is *Sturmian on $l$ letters* if $p(n) = n + l - 1$ for every $n \in \mathbb{N} \setminus \{0\}$. This is equivalent to $p(1) = l$ (and we may then suppose the alphabet $A$ has cardinality $l$) and $p(n+1) - p(n) = 1$ for any $n \geqslant 1$. A Sturmian sequence cannot be ultimately periodic, and the equation $p(n+1) - p(n) = 1$ is equivalent to the following property: there exists one word in $L_n(u)$ which is a prefix of two different words in $L_{n+1}(u)$, and each of the other words in $L_n(u)$ is a prefix of one and only one word in $L_{n+1}(u)$.

Let $k \geqslant 2$ be an integer; to any sequence $u$ on any alphabet $A$ included in $\{0, ..., k-1\}$ we associate the real number whose expansion in base $k$ is $0.u_0 u_1 \cdots u_n ...$, namely

$$S_k(u) = \sum_{n=0}^{+\infty} \frac{u_n}{k^{n+1}}.$$

*Throughout this paper, the base $k$ of any expansion is an integer.*

## 2. A COMBINATORIAL CRITERION FOR TRANSCENDENCE

PROPOSITION 1. *If $\theta$ is an irrational number and, for every $n \in \mathbb{N}$, the expansion of $\theta$ in base $k$ begins by $0.U_n V_n V_n V'_n$, where $U_n$ is a possibly empty and $V_n$ is a nonempty word on an alphabet $A \subset \{0, ..., k-1\}$, $V'_n$ is a prefix of $V_n$, $|V_n| \to +\infty$, $\limsup (|U_n|/|V_n|) < +\infty$ and $\liminf(|V'_n|/|V_n|) > 0$, then $\theta$ is a transcendental number.*

*Proof.* Let $r_n = |U_n|$, $s_n = |V_n|$, and choose $0 < \varepsilon < \liminf(|V'_n|/|V_n|)$. Let $t_n$ be the number whose expansion in base $k$ is $0.U_n V_n \cdots V_n \cdots$; then

$$t_n = \frac{p_n}{k^{r_n}(k^{s_n} - 1)}$$

for some integer $p_n$, while for $n$ large enough

$$|\theta - t_n| \leqslant \frac{1}{k^{r_n + (\varepsilon + 2) s_n}}.$$

Now, suppose $\theta$ were algebraic irrational. Then, from a theorem of Ridout ([MAH], Chap. 9, pp. 147–148), if there exist infinitely many rational numbers $P_n/Q_n$, with $Q_n = k^{m_n} Q'_n$ (the numbers $k$, $m_n$ and $Q'_n$ being integers), such that

$$\left| \frac{P_n}{Q_n} - \theta \right| < c_1 (Q_n)^{-\rho}$$

and

$$Q'_n < c_2 (Q_n)^{\mu},$$

where $c_1$ and $c_2$ are positive constants, then $\rho \leqslant 1 + \mu$. Since $\liminf(s_n/(r_n + s_n)) = 1/(\limsup(r_n/s_n) + 1) > 0$, and up to restricting $n$ to a strictly increasing sequence of integers, one can suppose that $s_n/(r_n + s_n) \to \eta > 0$. In particular, there exist two numbers $\rho$ and $\mu$ such that, for all $n$ in some infinite set,

$$1 + \frac{s_n}{r_n + s_n} < 1 + \mu < \rho < 1 + (1 + \varepsilon) \frac{s_n}{r_n + s_n}.$$

This choice of $\mu$ and $\rho$ together with the choice $P_n = p_n$, $Q_n = k^{r_n}(k^{s_n} - 1)$, $m_n = r_n$, and $Q'_n = k^{s_n} - 1$ gives us the desired contradiction. Hence $\theta$ is transcendental. Q.E.D.

## 3. THE STURMIAN CASE

Lemma 1. *If $u$ is Sturmian on $l$ letters and not recurrent, $u$ is ultimately equal to a Sturmian recurrent sequence on $l' < l$ letters.*

*Proof.* Suppose a word $W$, of length $m$, does not occur an infinite number of times in $u$; then there exists an $N$ such that the complexity of the sequence $v_n$ equal to $(u_n, n \geqslant N)$ satisfies $p(m) \leqslant m + l - 2$; but for each $n$, $L_n(v) \subset L_n(u)$, hence every word in $L_n(v)$ is a prefix of almost one word in $L_{n+1}(v)$, except maybe one which is a prefix of two words, and hence $p(n+1) - p(n) \leqslant 1$; but also $p(n+1) - p(n) > 0$ as this sequence is not ultimately periodic; hence it must have complexity $n + l_1 - 1$ for all $n \geqslant 1$, for some $0 \leqslant l_1 < l$; if it is recurrent, the lemma is proved; if it is not recurrent, we iterate the process, and, after at most $l$ steps, it shows that $u$ is ultimately equal to some recurrent Sturmian sequence. Q.E.D.

The same proof shows that every binary Sturmian sequence is recurrent; in fact, every binary Sturmian sequence, and, more generally every recurrent Sturmian sequence, can be shown to be minimal, for example by using the algebraic expressions given in Section 4.

The same method can be used ([ALE]) to prove that a non-recurrent Sturmian sequence $u$ is of the form $u_0 \cdots u_p v$, where $v$ is a recurrent Sturmian sequence on an alphabet $A'$, and $u_0, ..., u_p$ are distinct elements of $A \backslash A'$; this characterizes completely the non-recurrent Sturmian sequences.

The following lemma is essentially the same as in [ARN-RAU, pp. 205–208], though we need a different labelling of the edges in the graph and a different definition of what they call a segment; this yields correspondingly different, though of course equivalent, results.

LEMMA 2. *If $u$ is a recurrent Sturmian sequence, there exist two words $W_0$ and $W_1$, and a sequence of integers $a_n \geqslant 1$, $n \geqslant 1$, such that if the words $W_n$, $n \in \mathbb{N}$, are given by the recursion formulas*

$$W_{n+1} = W_n^{a_n} W_{n-1}$$

*for $n \geqslant 1$. Then, for any $N \geqslant 1$ and $n \geqslant 1$, the word $u_0 u_1 \cdots u_{N-1}$ is of the form $X_0 X_1 \cdots X_k$, where $X_1, X_2, ..., X_{k-1}$ are equal either to $W_n$ or to $W_{n+1}$, $X_0$ is a (possibly empty) suffix of either $W_n$ or $W_{n+1}$, $X_k$ is a (possibly empty) prefix of either $W_n$ or $W_{n+1}$. This decomposition, which is not unique, is independent of $N$ for fixed $n$.*

*Proof.* Let $A$ be the alphabet of the sequence. Let $L_n(u)$ be the set of all factors of $u$ of length $n$; let $\Gamma_n$ be the graph whose vertices are the elements of $L_n(u)$, and where there is an arrow from $E$ to $F$, with label $b$, whenever $E = bH$, $F = Ha$, with $b \in A$, $a \in A$, and $bHa \in L_{n+1}(u)$ (in [ARN-RAU] the same arrow is labelled with $a$). The property $p(n+1) - p(n) = 1$ implies there is one vertex $D_n$ with two outgoing arrows, and from every other vertex leaves only one arrow; the property $p(n+1) - p(n) = 1$ and the recurrence imply that there is one vertex $G_n$ with two incoming arrows, and to every other vertex arrives only one arrow; hence $\Gamma_n$ has one of the following two forms:

$$\begin{array}{ccccccc} & \leftarrow & & & & \leftarrow & \\ \downarrow & & \uparrow & & \downarrow & & \uparrow \\ G_n & \rightarrow & D_n & \text{or} & G_n & = & D_n. \\ \downarrow & & \uparrow & & \uparrow & & \downarrow \\ & \rightarrow & & & & \rightarrow & \end{array}$$

We call an *n-segment* any finite sequence $(E_0, ..., E_k)$ of vertices of $\Gamma_n$ such that $E_0 = G_n$, $E_i \to E_{i+1}$, $E_k = G_n$, and to each $E_i$, $1 \leqslant i \leqslant k-1$ arrives only one arrow. The *name* of an $n$-segment is the word made with the labels of the arrows $E_0 \to E_1, ..., E_{k-1} \to E_k$. There are exactly two $n$-segments for each $n$, and their names generate the language $L_n(u)$. Let $K_n$ and $J_n$ be the names of the two $n$-segments.

We follow now the reasoning of [ARN-RAU], to which we refer the reader for further details, with the necessary modifications (they use segments going from $D_n$ to $D_n$). There are two cases for going from $\Gamma_n$ to $\Gamma_{n+1}$: in the first one, $G_n \neq D_n$. Then for any word $X \neq D_n$ in $L_n(u)$, there exists a unique word $Xa$ in $L_{n+1}(u)$, and if $Yb \to Xa$, then $Y \to X$. We must have $G_{n+1} = G_n \gamma$ and $D_{n+1} = \delta D_n$, where $\gamma$ and $\delta$ are uniquely determined by the graph $\Gamma_n$. The graph $\Gamma_{n+1}$ is then known entirely, and we check that $K_{n+1} = K_n$ and $J_{n+1} = J_n$.

In the second case, $G_n = D_n$; let the two $n$-segments be $(G_n, G'_n a, ..., bG''_n, G_n)$ with name $K_n$, and $(G_n, G'_n c, ..., dG''_n, G_n)$, with name $J_n$. Then $G_{n+1} = G_n \gamma$ and $D_{n+1} = \delta G_n$ but now $\Gamma_n$ does not determine $\Gamma_{n+1}$. To lift the indetermination, we suppose for example that $\gamma = a$; then, because of the recurrence of $u$, we must have $D_{n+1} = bD_n$, the other possible choice giving a non-transitive graph. This implies that $K_{n+1} = K_n$, $J_{n+1} = K_n J_n$; the other choice for $\gamma$ leads to $K_{n+1} = J_n K_n$, $J_{n+1} = J_n$.

Hence the names of the two $n$-segments are $W_n$ and $W_{n+1}$, with the recursion formulas we claimed; and now, for any fixed $n$ and $N$, the word $u_0 \cdots u_{N-1}$ is the path in $\Gamma_n$ starting at $u_0 \cdots u_{n-1}$, continuing through $u_1 \cdots u_n$, $u_2 \cdots u_{n+1}$, ... and ending at $u_{N+1} \cdots u_{N+n}$. So it has the required decomposition. Q.E.D.

PROPOSITION 2. *If there exists k such that the expansion of $\theta$ in base k is a Sturmian sequence, then $\theta$ is a transcendental number.*

*Proof.* As the transcendence does not depend on the initial values of $u$, it is enough, because of Lemma 1, to prove our claim if $\theta = S_k(u)$ for a recurrent Sturmian sequence $u$. Let then $a_n$ and $W_n$ be as in Lemma 2.

Then, for each $n$, a suitable initial segment of $u$ is $X_0 X_1 \cdots X_{k-1}$ as in Lemma 2; $X_0$ is either a suffix of $W_n$, denoted by $T_n$, or a suffix of $W_{n+1}$, which may be a suffix of $W_{n-1}$, denoted again by $T_n$, or is of the form $T_n W_n^{c_n} W_{n-1}$ with $T_n$ a suffix of $W_n$ and $0 \leqslant c_n \leqslant a_n$ an integer (every considered suffix may be empty). Then the first $b_n$ words among $X_1, ..., X_{k-1}$ are $W_n$ for some integer $b_n \geqslant 0$, and then comes one $W_{n+1}$ (if not, $u$ would be ultimately periodic).

Hence, for every $n$, $u$ begins by either (1) the word $T_n W_n^{b_n + a_n} W_{n-1}$ or (2) the word $T_n W_n^{c_n} W_{n-1} W_n^{b_n + a_n} W_{n-1}$, where $T_n$ is a suffix of $W_n$ or of

$W_{n-1}$ and $b_n$ and $c_n$ are non-negative integers. Let $q_n$ be the length of $W_n$, satisfying $q_{n+1} = q_n a_n + q_{n-1}$. Then:

- if, for infinitely many $n$, the case (2) occurs with $c_n \geqslant 3$, Proposition 1 applied for this sequence with $U_n = T_n$ and $V_n = V'_n = W_n$ yields the transcendence of $S_k(u)$;

- if not, we take $U_n$ to be $T_n$ in case (1) and $T_n W_n^{c_n} W_{n-1}$ in case (2), so we have ultimately (i.e., for each $n$ large enough) $|U_n| \leqslant 5q_n$. And

— if $a_n + b_n \geqslant 3$ for infinitely many $n$, we take $V_n = W_n$ and apply Proposition 1, with $V'_n = V_n$, which yields the result;

— if $a_n + b_n \leqslant 2$ ultimately but $a_n + b_n = 2$ infinitely often, then $q_{n-1} \geqslant q_n/3$ ultimately, and Proposition 1 with $V_n = W_n$ and $V'_n = W_{n-1}$ yields the result;

— finally, in the remaining case we must have $b_n = 0$ and $a_n = 1$ ultimately. In this case, where the reader will recognize the Fibonacci recursion, we have also $W_n W_{n-1} = W_{n-1} W_{n-1} W_{n-4} W_{n-3}$, and we apply Proposition 1 with $V_n = W_{n-1}$ and $V'_n = W_{n-4}$, as $|V_n| = q_{n-1}$ is then larger than $|U_n|/10$ and smaller than $8\,|V'_n|$. Q.E.D.

For the particular cases in which the transcendence was already known, see the discussion after Lemma 3 below.

## 4. ALGEBRAIC EXPRESSION

We note that our method uses only the combinatorial properties of the Sturmian sequences, and that our Lemma 2, which gives only an explicit characterization of the *language* of $u$ but not of the sequence itself, is sufficient to yield the main result. There exist, however, more precise results than our Lemma 2, and they have been used to give the transcendence result in some particular cases.

There is a characterization, due to Morse and Hedlund ([HED-MOR]), of Sturmian sequences on $A = \{0, 1\}$: there exist an irrational number $\alpha$ and a real number $x$ such that, if $Ry = y + \alpha \bmod 1$, then either $u_n = 1$ whenever $R^n x \in [1-\alpha, 1[$, or $u_n = 1$ whenever $R^n x \in ]1-\alpha, 1]$. This gives an algebraic expression of S:

LEMMA 3. *If $u$ is a Sturmian sequence on $\{0, 1\}$, either*

$$S_k(u) = \sum_{n=1}^{+\infty} k^{-[(n-x)/\alpha]}$$

*or*

$$S_k(u) = \sum_{n=1}^{+\infty} k^{-[(n-x)/\alpha]'},$$

*with* $[x]' = [x]$ *if* $x$ *is not an integer, and* $[x]' = x+1$ *if* $x$ *is an integer.*

This algebraic expression allowed the transcendence of $S_k(u)$ to be proved for the homogeneous case, that is, $x=0$ ([ADA-DAV] or [BUL-SEN] for example; also [DAV] for the special subcase of the Fibonacci sequence), which corresponds to the case where (in the notations of Lemma 2) $u_0 \cdots u_{n-1} = G_n$ for every $n \geqslant 2$, and hence $u$ begins by the word $W_n$ for every $n$. Also, there have been several results of algebraic independence, in the homogeneous case, for similar expressions associated to different irrationals $\alpha$; see [ZHU] for example.

Furthermore, the sequence $u$ can be explicitly determined from the knowledge of $x$ and $\alpha$; however, the computations are very much involved in the nonhomogeneous case and the result in [NIS-SHI-TAM] had to be corrected in [KOM] (see also [ARN-FER-HUB] for a simpler expression of $u$); the $a_n$ of Lemma 2 are of course the coefficients of the continued fraction approximation of $\alpha$.

This explicit determination allowed transcendence and algebraic independence properties of $S_k(u)$ to be claimed in [NIS-SHI-TAM], for some particular cases intersecting the non-homogeneous case; [KOM] shows that part of these results are still true with the corrected expression of $u$: namely, when $\alpha$ has unbounded partial quotients in its simple continued fraction expansion, for all values of $x$, $S_k(u)$ is transcendental (and a Liouville number), and also some results of algebraic independence hold between different algebraic expressions associated to the same $\alpha$ and $x$. Another particular case of transcendence was claimed in [NIS-SHI-TAM] but could not be proved or disproved in [KOM]: it stated that the transcendence of $S_k(x)$ when some joint continued fraction expansion of $\alpha$ and $x$, which is usually known as the Ostrowski expansion ([OST]), stops after a finite number of steps; our Proposition 2 shows that this result is true, as is the transcendence of $S_k(u)$ for any irrational $\alpha$ and real $x$.

We shall now derive a similar algebraic expression for Sturmian sequences on $l$ letters:

LEMMA 4. *Let* $u$ *be a recurrent Sturmian sequence on an alphabet* $A$; *then there exist distinct elements* $e_1, ..., e_b, f_1, ..., f_c, g_1, ..., g_d$ *in* $A$ *such that the sets* $E = \{e_1, ..., e_b\}$, $F = \{f_1, ..., f_c\}$, $G = \{g_1, ..., g_d\}$ *are disjoint,* $E \cup F \cup G = A$, $G \neq \varnothing$, *and* $E \cup F \neq \varnothing$, *and there exists a Sturmian sequence* $v$ *on* $\{0, 1\}$ *such that, if* $\sigma$ *is the substitution* $0 \to g_1 \cdots g_d e_1 \cdots e_b$,

*$1 \to g_1 \cdots g_d f_1 \cdots f_c$, then $\sigma v = Wu$, where $W$ is a (possibly empty) prefix of $\sigma 0$ or $\sigma 1$.*

*Proof.* We apply Lemma 2: the language of $u$ is generated by words $W_n$, such that $W_{n+1} = W_n^{a_n} W_{n-1}$; $W_0$ and $W_1$ may be taken as the name of the two 1-segments in $\Gamma_1$; the definition of $\Gamma_1$ implies that $W_0 = g_1 \cdots g_d e_1 \cdots e_b$ and $W_1 = g_1 \cdots g_d f_1 \cdots f_c$ with the $e$, $f$, $g$ as in the conclusion of this lemma; the sequence $u$ itself may be written as an infinite concatenation $UW_{i_1} \cdots W_{i_n} \dots$, where $U$ is a suffix of $W_{i_0}$ with $i_0 = 0$ or $i_0 = 1$.

Let now $V_0 = 0$, $V_1 = 1$, and $V_{n+1} = V_n^{a_n} V_{n-1}$; then the sequence beginning by $V_n$ for every $n$ is Sturmian (and may be represented as in [HED-MOR] with $x = 0$); the sequence $v$ written as $i_0 V_{i_1} \cdots V_{i_n}$ is also Sturmian as it has the same language, and satisfies the conclusion. Q.E.D.

As can be seen from their graphs $\Gamma_n$, all sequences thus built are Sturmian, and therefore Lemma 4 together with the remark after Lemma 1 and the result of [HED-MOR] gives a complete characterization of all Sturmian sequences. We use it to compute $S_k(u)$:

PROPOSITION 3. *Let $u$ be a Sturmian sequence on an alphabet $A \subset \{0, \dots, k-1\}$. Then there exist distinct elements $e_1, \dots e_b$, $f_1, \dots, f_c$, $g_1, \dots, g_d$ in $A$ such that the sets $E = \{e_1, \dots, e_b\}$, $F = \{f_1, \dots, f_c\}$, and $G = \{g_1, \dots, g_d\}$ are disjoint $E \cup F \cup G = A$, $G \neq \emptyset$, and $E \cup F \neq \emptyset$; there exist a rational number $M$, an irrational number $\alpha$, and a real number $0 \leqslant x \leqslant 1$ such that, if $g_1 \cdots g_d e_1 \cdots e_b = e'_1 \cdots e'_r$, $g_1 \cdots g_d f_1 \cdots f_c = f'_1 \cdots f'_s$, then either*

$$S_k(u) = M + \sum_{t=1}^{r} \sum_{n \geqslant n_0(t)} \frac{e'_t}{k^{t+rn+s[x/(1-\alpha)+n\alpha/(1-\alpha)]}}$$

$$+ \sum_{t=1}^{s} \sum_{n \geqslant n_1(t)} \frac{f'_t}{k^{t+sn+r[-x/\alpha+n(1-\alpha)/\alpha]}}$$

*where $n_0(t)$ and $n_1(t)$ are the first values of $n$ for which the corresponding exponent of $k$ is positive, or*

$$S_k(u) = M + \sum_{t=1}^{r} \sum_{n \geqslant n_0(t)} \frac{e'_t}{k^{t+rn+s[x/(1-\alpha)+n\alpha/(1-\alpha)]'}}$$

$$+ \sum_{t=1}^{s} \sum_{n \geqslant n_1(t)} \frac{f'_t}{k^{t+sn+r[-x/\alpha+n(1-\alpha)/\alpha]'}},$$

*where $n_0(t)$ and $n_1(t)$ are the first values of $n$ for which the corresponding exponent of $k$ is positive.*

*Proof.* We apply Lemma 1 and Lemma 4 to get $S_k(u) = M + S_k(u')$, where $M$ is rational and $u' = \sigma v$ for some Sturmian sequence on $\{0, 1\}$ and some $\sigma$ as in Lemma 4; by [HED-MOR], we get $x$ and $\alpha$ such that, for example, if $Ry = y + \alpha \bmod 1$, then $v_n = 1$ whenever $R^n x \in [1-\alpha, 1[$ (if the interval is semi-open on the other side, we just replace the [ ] in the result by [ ]′).

Now, if we define the set $H \subset \mathbb{R}^2$ to be $\{(y, i),\ y \in [0, 1-\alpha[,\ 0 \leqslant i \leqslant r-1\} \cup \{(y, i),\ y \in [1-\alpha, 1[,\ 0 \leqslant i \leqslant s-1\}$ and the transformation $T$ on $H$ by $T(y, i) = (y, i+1)$ whenever $(y, i+1) \in H$, $T(y, i) = (Ry, 0)$ otherwise; then we get $u'_n = e'_i$ whenever $T^n(x, 0) = (y, i)$ and $y \in [0, 1-\alpha[$, and $u'_n = f'_i$ whenever $T^n(x, 0) = (y, i)$ and $y \in [1-\alpha, 1[$.

But $T$ is also the translation by $(0, 1)$ in $\mathbb{R}^2$ modulo the lattice $L$ whose $H$ is a fundamental domain, that is, the lattice generated, for example, by the vectors $(-\alpha, r)$ and $(1-\alpha, s)$. Hence, we can write that $u'_p = e'_i$ if and only if there exist integers $l$ and $m$ and a point $y \in [0, 1-\alpha[$ such that $x - y = -l\alpha + m(1-\alpha)$ and $p - i = rl + sm$, and this allows us to identify the values of $p$ for which $u'_p = e'_i$; a similar identification for $u'_p = f'_j$ finally yields the desired expression. Q.E.D.

## 5. GENERALIZATION TO ARNOUX–RAUZY SEQUENCES

Some sequences share part of the combinatorial properties of the Sturmian sequences: these are the sequences of complexity $p(n) = 2n+1$, with an additional property called the "star condition" in [ARN-RAU]. We refer the reader to that paper for the proof of Lemma 5, into which we incorporated the same modifications, in the labelling of edges and the definition of segments, as in Lemma 2.

Definition. A sequence $(u_n, n \in \mathbb{N})$ on an alphabet $A$ with three letters is an *Arnoux–Rauzy* sequence if it has the following four properties:

- it is minimal;
- its complexity satisfies $p(n) = 2n+1$ for every $n \geqslant 1$;
- every word in $L_n(u)$ is a prefix of exactly one word in $L_{n+1}(u)$, except one which is a prefix of three words;
- every word in $L_n(u)$ is a suffix of exactly one word in $L_{n+1}(u)$, except one which is a suffix of three words.

Lemma 5. *If $u$ is an Arnoux–Rauzy sequence, then there exist three words, $A_0$, $B_0$, and $C_0$, and a sequence of integers, $1 \leqslant i_n \leqslant 3$, $n \geqslant 1$, taking infinitely often the values* 1, 2, *and* 3, *such that if the words $A_n$, $B_n$, and $C_n$, $n \in \mathbb{N}$, are given by the recursion formulas*

$$A_{n+1} = A_n$$
$$B_{n+1} = A_n B_n$$
$$C_{n+1} = A_n C_n,$$

*if* $i_n = 1$;

$$A_{n+1} = B_n A_n$$
$$B_{n+1} = B_n$$
$$C_{n+1} = B_n C_n,$$

*of* $i_n = 2$; *and*

$$A_{n+1} = C_n A_n$$
$$B_{n+1} = C_n B_n$$
$$C_{n+1} = C_n$$

*if* $i_n = 3$; *then for any* $N \geqslant 1$ *and* $n \geqslant 1$, *the word* $u_0 u_1 \cdots u_{N-1}$ *is of the form* $X_0 X_1 \cdots X_k$, *where* $X_1, X_2, ..., X_{k-1}$ *are equal to* $A_n$, $B_n$, *or* $C_n$; $X_0$ *is a* (*possibly empty*) *suffix of* $A_n$, $B_n$, *or* $C_n$; *and* $X_k$ *is a* (*possibly empty*) *prefix of* $A_n$, $B_n$, *or* $C_n$. *This decomposition, which is not unique, is independent of* $N$ *for fixed* $n$.

PROPOSITION 4. *If there exists* $k$ *such that the expansion of* $\theta$ *in base* $k$ *is an Arnoux–Rauzy sequence, then* $\theta$ *is a transcendental number.*

*Proof.* Let $u_n$ be an Arnoux–Rauzy sequence, and let $i_n$ be the sequence of 1, 2, and 3 giving its recursion formulas as in Lemma 5; let $\theta = S_k(u)$.

Throughout this proof, we say that *we are in the situation of Proposition* 1 if the expansion of the number $\theta$ in base $k$ begins with $0.UVVV'$, where $U$ is a possibly empty word and $V$ is a nonempty word on an alphabet $A \subset \{0, ..., k-1\}$; $V'$ is a prefix of $V$, $|U|/|V| < 100$ and $|V|/|V'| < 100$. If we are in the situation of Proposition 1 for an infinite number of different words $V$, then $\theta$ is transcendental.

We first suppose that the *sequence* $i_n$ *contains an infinite number of words* $aaa$, *for* $a = 1, 2,$ *or* 3.

If we see, for example, the word 111 in $(i_n)$, we claim that we are at least once in the situation of Proposition 1, in such a way that an infinite number of 111 will put us infinitely often in this situation:

If we see 111, then, for some $m > n$, we have $A_m = A_n$, $B_m = A_n^g B_n$, $C_m = A_n^g C_n$, with $g \geqslant 3$ and $i_m \neq 1$. Hence, by Lemma 6, the sequence $u$ begins by $T_n A_n^3$, where $T_n$ is a suffix of $A_n^g B_n$, or $A_n^g C_n$, or $A_n^g B_n A_n$, or

$A_n^g C_n A_n$, or $A_n^g B_n A_n^g C_n$, or $A_n^g C_n A_n^g B_n$. Hence we are in the situation of Proposition 1, with $V = V' = A_n$ and a suitable $U$, except in two cases: the case where $T_n$ is a strict suffix of $A_n^3 B_n$, as the length of $B_n$ might be large compared to the length of $A_n$—we say in this case that we have a 111 word with *direction* 2—and the case where $T_n$ is a strict suffix of $A_n^3 C_n$, where we say we have a 111 word with direction 3.

Suppose for example we have a 111 word with direction 2, and let $A$, $B$, $C$ be the three $n$-words of the last paragraph; we look in the sequence $(i_n)$ at the last occurrence of 2 before the word 111, and suppose it occurs in $(i_n)$ $f$ digits before the word 111; after the recursion step corresponding to this 2, the lengths of the three words $a$, $b$, $c$ satisfy $|b| < |a|$ and $|b| < |c|$; then, there are only recursion rules 1 and 3, and after $f$ of these rules we have $|B| < f\,|A|$ and $|B| < f\,|C|$.

We shall prove now that either we are in the situation of Proposition 1 at some place between this 111 and the last 2 before it, or, possibly after changing our 111 for an earlier occurrence of 111 or 333, we can replace $f$ by a fixed number (namely 24) in this bound.

To achieve that, we look more closely at the sequence $(i_n)$ between our 2 and our 111; it is, for example, of the form $1^{a_0}3^{a_1}\cdots$. Suppose there are others 111 in this part of the sequence $(i_n)$. For any one of them, we are in the situation of Proposition 1 if it does not have direction 2 or 3; if it has direction 2, we say it is *usable*; if it has direction 3, we say it is usable whenever it does not lie in the initial $1^{a_0}$. If we find a 333 in this part of the sequence $(i_n)$, either we are in the situation of Proposition 1 or, whether the 333 has direction 2 or 1, we call it usable. The symmetrical case applies if the sequence $(i_n)$ between our originals 2 and 111 is of the form $3^{a_0}1^{a_1}\cdots$. Then, between our original 111 and the last 2 before it, either we are at least once in the situation of Proposition 1 and our claim is proved, or there are usable words 111 or 333. In the second case, let $xxx$ be the first usable word occurring in $(i_n)$ at the right of the considered 2; we call it the *useful* $xxx$ word, and let $y$ be its direction. Hence, if we look at the useful $xxx$ with direction $y$ (instead of our original 111 with direction 2), we know that between $xxx$ and the last $y$ before it in $(i_n)$, there are only digits $z \neq y$, with no $zzz$ except maybe in the initial word of identical digits after $y$. Furthermore, the useful $xxx$ occurs in $(i_n)$ at a place situated between our original 111 and the last digit 2 before, hence an infinite number of original 111 will still yield an infinite number of useful $xxx$.

So, for example, suppose the useful $xxx$ is still an 111 with direction 2 and the sequence $(i_n)$ between the last 2 and the useful 111 is $1^M 3^{a_1} 1^{a_2}\ldots$, with $a_1 \leqslant 2$, $a_2 \leqslant 2$, ..., but maybe $M$ is large.

Suppose that $M \geqslant 3$; after the initial $1^M$ and the following 3, the words are $a^M ca$, $a^M ca^M b$, and $a^M c$, and either we are in the situation of Proposition 1 or we have a 111 with direction 3, and $u$ begins by a strict suffix of

$a^3c$. If $|c| \leqslant M\,|a|$, we are in the situation of Proposition 1 with $V = V' = a^{[M/3]}$; if $|c| > M\,|a|$, then $|a^M c a^M b| \leqslant 2|a^M ca|$. This means that, if $M \geqslant 3$, as well as directly if $M \leqslant 2$, we have, when we arrive at the useful 111, $|B| < 3(f - M)|A|$: so, by replacing $f$ by $3f$ in our bound on $|B|$, we can just forget the initial segment $1^M$. Note that if the initial segment is a $3^M$, after it the $b$ word will remain smaller than the $a$ word, and the length of $B$ is bounded in that case also.

Now, in the remaining $3^{a_1}1^{a_2}$..., suppose we find 131313; then we check that there exist words $a'$, $b'$, $c'$ such that $u$ begins by $Ta'c'a'a'c'a'a'c'$; hence we are in the situation of Proposition 1 with $V = a'c'a'$, $V' = a'c'$, and $U = T$ except if $T$ contains a suffix of $b'$. If the 131313 occurs in $(i_n)$ less than 10 digits after the word $21^M$, then $b'$ will not be longer than 20 times the length of $a'$, and we are in the situation of Proposition 1 even when $T$ contains a suffix of $b'$.

A similar reasoning applies if any one of the words 13311, 1311, 3133, 31133, and 313131 occurs in $(i_n)$ less than 10 digits after the $21^M$. But then we check the possible words that may appear after the $21^M$ and see that if we are never in the situation of Proposition 1, then our useful 111 must occur in $(i_n)$ not more than eight digits after the $21^M$, and hence $|B| < 24\,|A|$, and finally we are in the situation of Proposition 1 for the word $V$ coming from the useful 111.

Hence, through all these cases, we have proved that, because of the infinite number of 111, we are in the situation of Proposition 1 for an infinite number of different words $V$, and the number $\theta$ is transcendental.

*So now we can suppose that ultimately the sequence* $(i_n)$ *has no words* 111, 222, *or* 333.

We shall show now that we can eliminate the words $aa$; suppose that we see 11 in $(i_n)$, but not 111 of course. Then $u$ will begin by either $T_nA_nA_nB_n$ or $T_nA_nA_nC_n$. The possible problem of the length of $T_n$ is solved in the same way as before: if for example $T_n$ contains a suffix of $B_n$, we look at the last two in the sequence $(i_n)$, and show that either $|B_n| < 24\,|A_n|$ or we are at least once in the situation of Proposition 1 for some word $V$ which is a concatenation of a bounded number of words $A_m$, $B_m$, or $C_m$, with an $m$ lying between the place where our 11 occurs in $(i_n)$ and the place of the last occurrence of 2 before it (of course, as we now know that there are no 111 and 333, we only need to repeat the arguments about words 131313, 13311, and the like); so we can take $U = T_n$.

So, let us suppose for example that $u$ begins by $TA_nA_nB_n$; then $A_n = B_{n-1}A_{n-1}$, $B_n = B_{n-1}$ (resp., $A_n = C_{n-1}A_{n-1}$, $B_n = C_{n-1}B_{n-1}$). The now usual reasoning shows that either we are at least once in the situation of Proposition 1, for some word $A_p$, $B_p$, or $C_p$ with a $p$ lying between the place where our 11 occurs in $(i_n)$ and the place of the last occurrence of 2 (resp. 3) before it, or we have $|B_{n-1}| > |A_{n-1}|/24$ (resp.,

$|C_{n-1}| > |A_{n-1}|/24$). Hence in these last cases we are in the situation of Proposition 1 with $V = A_n$ and $V' = B_{n-1}$ (resp. $V' = C_{n-1}$).

An analogous reasoning takes care of the remaining cases; if we know 11, 22, and 33 will not appear ultimately, the words 121 and 1231 themselves allow us to be in the situation of Proposition 1. Hence in every case we are in the situation of Proposition 1 for infinitely many different words $V$ and $\theta$ is transcendental. Q.E.D.

## 6. OTHER GENERALIZATIONS AND PROBLEMS

Our Proposition 1 is well suited to words generated by primitive substitutions, which are known to be of sub-affine complexity ([PAN] using the results in [EHR-LEE-ROZ], or [COB] in the particular case of constant length; see [QUE] for a short proof):

PROPOSITION 5. *If the expansion of $\theta$ in some base $k$ is a non-ultimately periodic fixed point $u$ of a primitive substitution $\sigma$, and does contain at least one word of the form $V^{2+\varepsilon}$ (that is, $VVV'$ for a nonempty word $V$ and a prefix $V'$ of $V$ with $|V'| \geqslant \varepsilon\,|V|$), then $\theta$ is transcendental.*

*Proof.* We just write $u = UVVV'v$, for some finite word $U$ and some sequence $v$, and apply Proposition 1 with $U_n = \sigma^n U$, $V_n = \sigma^n V$, and $V'_n = \sigma^n V'$ for every $n \geqslant 0$; it is well known (see for example [QUE]) that there exists a real number $\lambda$ such that

$$\frac{|\sigma^n W|}{\lambda^n\,|W|} \to c(W) > 0$$

when $n \to +\infty$ for every word $W$, and this gives the required conditions on the lengths. Q.E.D.

When $\sigma$ is of constant length, such an $u$ is an automatic sequence, and hence we are in a particular case of the theorem of Loxton and van der Poorten [LOX-vdP], but with a straightforward proof (and the referee states that the paper of Loxton and van der Poorten has been reported to contain a gap).

The following result generalizes Theorem 3 of [TAM]:

COROLLARY 1. *Let $\omega$ be a fixed point of a primitive substitution $\sigma$ on the alphabet $\{a_1, ..., a_s\}$, containing at least one word of the form $V^{2+\varepsilon}$, and let $\tau$ be any substitution from $\{a_1, ..., a_s\}$ to $\{0, ..., g-1\}$; if $\tau\omega$ is not ultimately periodic and is the expansion in base $g$ of $\theta$, then $\theta$ is transcendental.*

It is a corollary of the proof of Proposition 5: we can replace $\sigma^n V$ by $\tau\sigma^n V$, and the same for $U$ and $V'$, the ratios in the hypotheses of Proposition 1 being just multiplied or divided by the maximum length of the $\tau a_i$.

The particular case studied in [TAM] is the case where $\sigma(a_j) = a_1^{k_j} a_{j+1}$ for $1 \leqslant j \leqslant s-1$, $\sigma(a_s) = a_1$, for integers $1 \leqslant k_1 \leqslant \cdots \leqslant k_{s-1}$; the class of sequences thus studied has a nonempty intersection with the homogeneous Sturmian sequences, for example, with the Fibonacci substitution and the Arnoux–Rauzy sequences, but of course not all Sturmian or Arnoux–Rauzy sequences are generated by one substitution in this way; it includes also sequences of higher complexity such as $sn+1$ for any natural integer $s$. The fixed point $\omega$ of $\sigma$ satisfies the hypotheses of our Corollary 1 as it must contain $a_s a_1$, and hence either $a_1 a_1 a_1$ or $a_1 a_1 a_2$, and hence $a_1^{k_1} a_2 a_1^{k_1} a_2 a_1$. In that paper, Tamura also shows the linear independence over $\mathbb{Q}$ of numbers associated to different substitutions $\tau$ applied to one fixed point $\omega$.

Let us remark though that our Propositions 1 and 5 do not apply to the famous Thue–Morse substitution $\rho 0 = 01$, $\rho 1 = 10$: its fixed points are known ([THU]) to contain no factor of the form $wwx$ where $x$ is the first letter of $w$, though it is well known that $\sum_n u_n 2^{-n}$ is transcendental when $u$ is the Thue–Morse sequence on $\{0, 1\}$ or $\{-1, +1\}$ ([DEK]).

Another example where Proposition 1 applies is Chacon's sequence, which is the fixed point beginning with 0 of the (non-primitive) substitution $\tau 0 = 0010$, $\tau 1 = 1$ (take for example $U_n = \tau^n 001$, $V_n = V' = \tau^n 0$); this sequence has complexity $2n-1$ for $n \geqslant 2$ ([FER]). Both these results are still true for any sequence with the same language.

The transcendence question for general sequences of complexity $2n+1$ is an open problem. Another still open question is to find a geometric representation for the Arnoux–Rauzy sequences, involving rotations in $\mathbb{R}^2$; this would give an algebraic expression for the associated transcendental numbers, as in Section 5.

As was noticed in [ALL-ZAM], it follows from a result in [BER-SEE] that, except for the Thue–Morse substitution, any non-periodic fixed point of a substitution on two letters which is either primitive or of constant length satisfies the hypotheses of our Proposition 1.

## REFERENCES

[ADA-DAV] W. W. Adams and J. L. Davison, A remarkable class of continued fractions, *Proc. Amer. Math. Soc.* **65** (1977), 194–198.

[ALE] P. Alessandri, "Codages de rotations et basses complexités," Ph.D. thesis, University Aix-Marseille 2, 1996.

[ALL-ZAM] J.-P. Allouche and L. Q. Zambon, Algebraic irrational binary numbers cannot be fixed points of non-trivial constant length or primitive substitutions, preprint.

[ARN-FER-HUB] P. Arnoux, S. Ferenczi, and P. Hubert, Trajectories of rotations, Preprint 97-08, Institut de Mathématiques de Luminy, Marseille, 1997.

[ARN-RAU] P. Arnoux and G. Rauzy, Représentation géométrique de suites de complexité $2n+1$, *Bull. Soc. Math. France* **119** (1991), 199–215.

[BER-SEE] J. Berstel and P. Séébold, A characterization of overlap-free morphisms, *Disc. Appl. Math.* **46** (1993), 275–281.

[BUL-SEN] S. Bullett and P. Sentenac, Ordered orbits of the shift, square roots, and the devil's staircase, *Math. Proc. Camb. Phil. Soc.* **115** (1994), 451–481.

[COB] A. Cobham, Uniform tag sequences, *Math. Systems Theory* **6** (1972), 164–192.

[DAV] J. L. Davison, A series and its associated continued fraction, *Proc. Amer. Math. Soc.* **63** (1977), 29–32.

[DEK] M. Dekking, Transcendance du nombre de Thue–Morse, *C. R. Acad. Sci. Paris Sér. A-B* **285** (4) (1977), A157–A160.

[EHR-LEE-ROZ] A. Ehrenfeucht, K. P. Lee, and G. Rozenberg, Subword complexities of various classes of deterministic developmental languages without interactions, *Theoret. Comput. Sci.* **1** (1975), 59–75.

[FER] S. Ferenczi, Les transformations de Chacon: combinatoire, structure géométrique, liens avec les systèmes de complexité $2n+1$, *Bull. Soc. Math. France* **123** (1995), 271–292.

[HED-MOR] G. A. Hedlund and M. Morse, Symbolic dynamics II: Sturmian trajectories, *Amer. J. Math.* **62** (1940), 1–42.

[KOM] T. Komatsu, A certain power series and the inhomogeneous continued fraction expansions, *J. Number Theory* **59** (1996), 291–312.

[LOX-vdP] J. H. Loxton and A. J. van der Poorten, Arithmetic properties of automata: regular sequences, *J. Reine Angew. Math.* **392** (1988), 57–69.

[MAH] K. Mahler, Lectures on diophantine approximations. Part 1. *g*-adic numbers and Roth's theorem, University of Notre Dame, 1961.

[NIS-SHI-TAM] K. Nishioka, J.-i. Tamura, and I. Shiokawa, Arithmetical properties of a certain power series, *J. Number Theory* **42** (1992), 61–87.

[OST] A. Ostrowski, Bemerkungen zur Theorie der Diophantischen Approximationen I, II, *Abh. Math. Sem. Hamburg* **1** (1922), 77–98, 250–251.

[PAN] J.-J. Pansiot, Complexité des facteurs des mots infinis enegendrés par morphismes itérés, *in* "ICALP '84," Lecture Notes in Computer Science, Vol. 172, pp. 380–389, Springer-Verlag, New York/Berlin, 1984.

[QUE] M. Queffélec, "Substitution Dynamical Systems–Spectral Analysis," Lecture Notes in Math., Vol. 1294, Springer-Verlag, New York/Berlin, 1987.

[TAM] J.-i. Tamura, A class of transcendental numbers having explicit *g*-adic and Jacobi–Perron expansions of arbitrary dimension, *Acta. Arith.* **71** (1995), 301–329.

[THU] A. Thue, Über die gegenseite Lage gleiche Teile gewisser Zeichenreihen, *Norske vi. Selsk. Skr. Mat. Nat. Kl.* **1** (1912), 1–67; reprinted in "Selected Mathematical Papers of Axel Thue" (T. Nagell, Ed.), pp. 413–478, Universitetsforlaget, Oslo, 1977.

[ZHU] Y. Zhu, The algebraic independence of certain transcendental continued fractions, *Acta. Math. Sin.* **2** (1991), 127–134.